Journal of Cheminformatics

POSTER PRESENTATION Open Access

Hit series selection in noisy HTS data: clustering techniques, statistical tests and data visualisations

Christoph Müller[1*], Daniel Ormsby[1], Isabella Feierberg[2], Ola Engkvist[2], Christian Tyrchan[2], Michael J Hartshorn[1]

From 9th German Conference on Chemoinformatics
Fulda, Germany. 10-12 November 2013

High throughput screening (HTS) is one of the most prominent techniques used in the beginning stages of a drug discovery programme to identify those few hit compounds that can be used as starting points in subsequent studies [1,2]. However, an HTS experiment often entails a very data-intensive and challenging hit prioritization process that yields the mentioned hit compounds. The workflow described in this study aims to make this decision-making process easier by combining the structural and biological information of compounds used in an HTS. In particular, the workflow combines various clustering and nearest neighbourhood schemes with a non-parametric statistical test in order to prioritize those groupings of compounds that are likely of being relevant to the biological target of interest [3].

The novel workflow was evaluated under various aspects in a retrospective study using publicly available quantitative HTS (qHTS) datasets [4]. One of the main benchmarking aspects in this study was the ability to correctly identify as many true active compounds as possible. Therefore different chemical descriptors and clustering schemes were tested in combination with the statistic to measure their classification performance.

The workflow was integrated into Dotmatics' *Vortex*, a platform for analysing chemical information using chemoinformatics methods and data visualisations tools [5]. This integration enables researchers to easily extend their current HTS workflow in order to discover new hit series and reveal hidden relationships between compounds, scaffolds and clusters.

Authors' details
[1]Dotmatics Ltd, Windhill, Bishop's Stortford, CM23 2ND, UK. [2]AstraZeneca AB, Peppardsleden 1, Mölndal, 43183, Sweden.

* Correspondence: christoph.mueller@dotmatics.com
[1]Dotmatics Ltd, Windhill, Bishop's Stortford, CM23 2ND, UK
Full list of author information is available at the end of the article

Published: 11 March 2014

References

1. Rocke D: **Design and analysis of experiments with high throughput biological assay data.** *Cell and Developmental Biology* 2004, **15(6)**:703-713.
2. Keseru GM, Makara GM: **Hit discovery and hit-to-lead approaches.** *Drug Discovery Today* 2006, **11(15-16)**:741-748.
3. Varin T, Gubler H, Parker CN, Zhang JH, Raman P, Ertl P, Schuffenhauer A: **Compound set enrichment: A novel approach to analysis of primary HTS data.** *J Chem Inf Model* 2010, **50(12)**:2067-2078.
4. [http://www.ncbi.nlm.nih.gov].
5. Dotmatics Ltd:[http://www.dotmatics.com].

doi:10.1186/1758-2946-6-S1-P27
Cite this article as: Müller *et al.*: Hit series selection in noisy HTS data: clustering techniques, statistical tests and data visualisations. *Journal of Cheminformatics* 2014 **6**(Suppl 1):P27.